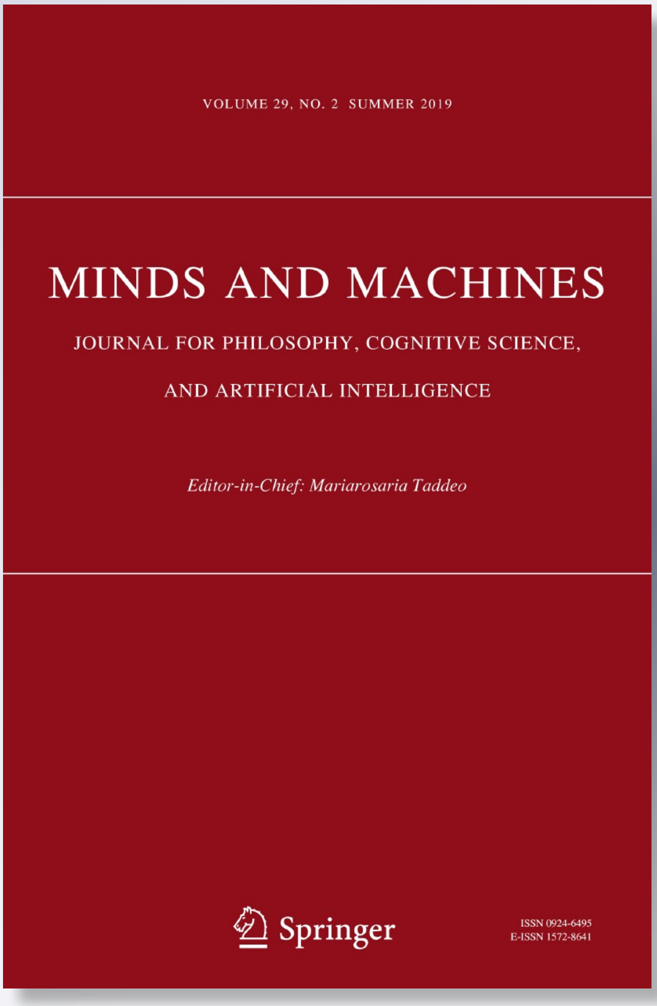
Triviality Arguments Reconsidered

Paul Schweizer

Minds and Machines
Journal for Artificial Intelligence,
Philosophy and Cognitive Science

ISSN 0924-6495
Volume 29
Number 2

Minds & Machines (2019) 29:287-308
DOI 10.1007/s11023-019-09501-x



VOLUME 29, NO. 2 SUMMER 2019

MINDS AND MACHINES

JOURNAL FOR PHILOSOPHY, COGNITIVE SCIENCE,
AND ARTIFICIAL INTELLIGENCE

Editor-in-Chief: Mariarosaria Taddeo

Springer

ISSN 0924-6495
E-ISSN 1572-8641

Springer

Minds and Machines (2019) 29:287–308
https://doi.org/10.1007/s11023-019-09501-x

ORIGINAL ARTICLE

Triviality Arguments Reconsidered

Paul Schweizer[1]

Received: 15 October 2018 / Accepted: 21 May 2019 / Published online: 28 May 2019

Abstract
Opponents of the computational theory of mind (CTM) have held that the theory is devoid of explanatory content, since whatever computational procedures are said to account for our cognitive attributes will also be realized by a host of other 'deviant' physical systems, such as buckets of water and possibly even stones. Such 'triviality' claims rely on a simple mapping account (SMA) of physical implementation. Hence defenders of CTM traditionally attempt to block the trivialization critique by advocating additional constraints on the implementation relation. However, instead of attempting to 'save' CTM by constraining the account of physical implementation, I argue that the general form of the triviality argument is *invalid*. I provide a counterexample scenario, and show that SMA is in fact consistent with empirically rich and theoretically plausible versions of CTM. This move requires rejection of the computational sufficiency thesis, which I argue is scientifically unjustified in any case. By shifting the 'burden of explanatory force' away from the concept of physical implementation, and instead placing it on salient aspects of the target phenomenon to be explained, it's possible to retain a maximally liberal and unfettered view of physical implementation, and at the same time defuse the triviality arguments that have motivated defenders of CTM to impose various theory-laden constraints on SMA.

Keywords Computational theory of mind · Computational sufficiency thesis · Physical implementation · Simple mapping account

1 Introduction

According to the long standing and widely embraced (e.g. Putnam 1967; Fodor 1975; Newell and Simon 1976; Stich 1983; Pylyshyn 1984; Johnson-Laird 1988; Pinker 1997) computational theory of mind (CTM), computation (of one sort or another) is held to provide the *scientific* key to explaining the mind, and in

✉ Paul Schweizer
paul@inf.ed.ac.uk

[1] Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh, Dugald Stewart Building, 3 Charles Street, Edinburgh EH8 9AD, UK

principle at least, computation provides the basis for the *engineering project* of constructing mental systems artificially. Actual physical devices such as laptops and brains perform computations when they implement abstract mathematical formalisms. And this immediately raises a question central to the foundations of CTM—under exactly what conditions can a physical system properly be said to implement a formal procedure? The answer to this question has proved surprisingly elusive and controversial.

A very straightforward and elegant account articulated by, e.g. Block (1978), Searle (1980), Hinkfuss (as reported in Lycan 1981), Kripke (1982), Putnam (1988), Bishop (2009), is based on a simple mapping between physical structure and abstract formalism. Accordingly, a physical system *P* performs a computation *C* just in case there is a *mapping* from the actual physical states of *P* to the abstract computational states of *C*, such that the transitions between physical states reflect the abstract state transitions as specified by the mapping. The minimalism, neutrality and generality of the simple mapping account (henceforth SMA, adopting terminology introduced in Godfrey-Smith 2009 and 'canonized' by Piccinini 2015a) make it *look* like a natural choice as the in-principle standard for physical implementation—it takes the Mathematical Theory of Computation (MTC) as its starting point and adds no substantial assumptions. And because it adds no further assumptions or restrictions, SMA is in an important sense maximally liberal—there will exist abstract mappings from a *huge* class of physical systems and processes to an equally huge class of computational formalisms.

Abstract mathematical structures are multiply realizable (MR), and SMA allows for the widest possible reading of this multiplicity. Indeed, one of CTM's standardly conceived virtues is that it can potentially provide a *universal* theory of cognition, a theory which is not tethered to the peculiarities and contingencies of biologically bestowed human physiology, and hence is free of the looming stigma of 'neurochauvinism'. Since mentality is explained in computational terms, and computational formalisms are multiply realizable, it follows that CTM can be applied to any number of different types of creatures and agents. In light of CTM in combination with MR, it follows that a human, a Martian and a robot could all be in exactly the *same* mental state, where this sameness is captured in terms of implementing the same cognitive computation, albeit via radically different forms of physical hardware. So on this view, computation is seen as providing the scientific paradigm for explaining mentality in general.

But rather than heralding multiple realizability as a theoretical virtue promising a universal account of mentality, various opponents of CTM target this feature as its Achilles heel. A maximally liberal reading of MR has been utilized by critics such as Searle (1980) and Putnam (1988), who contend that the theory is thereby rendered empirically vacuous. These 'trivialization' arguments hold that, *a la* SMA, a mapping will obtain between virtually any physical system and virtually any formalism, which in turn is construed as fatally undermining CTM, since whatever computational procedures are held to account for our cognitive properties and attributes will also be realized by a myriad of other 'deviant' physical systems. So, not only could human mental states be formally sustained in robots and extraterrestrials, but also by buckets of water and possibly even stones. Hence the attribution of computational structure to physical systems is deemed conceptually trivial, and devoid of any significant empirical content.

This triviality critique is certainly not aimed at a Philosophical Straw Man, since a host of authors have responded, including Fodor (1981), Maudlin (1989), Chrisley (1994), Chalmers (1996), Copeland (1996), Scheutz (1999), Shagrir (2001), Godfrey-Smith (2009), Sprevak (2010), Milkowski (2013), Rescorla (2014), Piccinini (2015b), in an attempt to defend CTM. These defences generally accept the basic supposition that an unbridled simple mapping account critically undermines the computational theory of mind, and hence they attempt to defend CTM by placing additional constraints on the implementation relation, so that it is no longer a simple or theoretically neutral mapping. In effect, these restrictions aim to salvage CTM by precluding a vast number of physical systems from the domain of the mapping function, in an attempt to separate 'true' or 'genuine' implementations from the many presumably 'false' cases countenanced by SMA. These constraints include: counterfactual, semantic, causal, functional, and mechanistic.

However, I advance quite a different type of response to the situation. Instead of attempting to 'save' CTM by constraining the account of physical implementation, I argue that, regardless of whether one accepts or rejects SMA, the general form of the triviality argument is *invalid*. I provide a counterexample scenario, and show that SMA is in fact *consistent* with empirically rich and scientifically plausible versions of CTM. Indeed, no one thinks that SMA 'threatens' electrical engineering or our ability to design and utilize sophisticated computational artefacts—and the Mathematical Theory of Computation is surely not seen as a source of peril by computer scientists and software engineers. So why should a purely MTC based account of physical implementation constitute such a theoretical hazard for a computational approach to the mind? In my view, it's a threat only because the widely embraced *version* of CTM which is undermined by SMA is itself flawed and should not be accepted in any case.

So I will argue that it is possible to retain SMA as the in-principle standard for physical implementation, while at the same time embracing a robust and empirically plausible rendition of CTM which is not susceptible to the standard trivialization critiques. The paper will proceed by first analyzing the trivialization strategy and arguing that the underlying Computational Sufficiency Thesis (CST) should be rejected. I then develop a modified version of CTM that is compatible with SMA, and show that this version of CTM could yield an explanatorily rich and scientifically fruitful account of the mind. Hence the triviality critique is invalidated. I then go on to provide some independent considerations in favor of SMA as the global, in-principle standard for physical implementation, and argue that those who accept CST and hence propose various additional constraints on SMA have inverted the picture, and that the 'burden of explanatory force' should not be borne by the implementation relation. I also propose that a genuinely *computational* account of mentality should include a specification of the formal abstract state transitions mediating inputs and outputs.

2 SMA and Trivialization

Putnam (1988) illustrates the inherent liberality of the simple mapping account with a technical proof of the theorem that every open physical system implements every (inputless) Finite State Machine (FSM). He provides a generic depiction of a physical system as a bounded, continuous region of space–time, and the basic idea is that the region is held constant but sliced up in an as many different ways as one likes in order to define a sequence of disjunctive 'physical states' that can be mapped to any given run of a FSM. And Searle famously promulgates the universality of SMA with the claim that virtually *any* physical system can be interpreted as implementing virtually *any* formal procedure. For example, Searle (1990) asserts that the molecules in his wall could be interpreted as running the WordStar program. The claim is simply put forward with no further defense, but Copeland (1996) provides a proof of 'Searle's Theorem', which he observes is essentially a notational variant of Newman's (1928) objection to Russell.

In light of such results, we will accept the crucial 'trivialization' premise that, given SMA, whatever computational procedures are held to account for our cognitive attributes will also be realized by a myriad of other 'nonstandard' physical systems. And merely for the sake of convenience, let us suppose the former to be some suitably advanced version of Fodor's (1975) Language of Thought (LOT), say LOT*. Hence assume that cognitive scientists eventually endorse LOT* as the underlying functional/computational architecture of the human mind, and thus hold that the brain serves as a biological implementation of this formal structure. And by SMA, a mapping will exist from the molecular activity in, say, Hinkfuss's pail, to this very same formal structure, so that the bucket of water also has a level of description under which it serves as an implementation of LOT*.

Given CTM, does it now follow that Hinkfuss's pail is on a cognitive par with the human mind and the trivialization strategy has succeeded? Not without a tacit underlying premise in the form of what Chalmers (2012) has dubbed the computational sufficiency thesis (CST). The CST explicitly maintains that merely implementing a computational formalism of the appropriate sort constitutes a *sufficient condition* for mentality in physical systems. So in order to diagnose the exact sense in which SMA is supposed to undermine CTM, it is useful to make the full structure of the trivialization strategy explicit with the following *reductio* argument:

Premise (1) Common sense pre-theoretical truth: this bucket of water *is not* a mental system

Premise (2) CTM (generic version): implementation of the appropriate formal architecture is fundamental to the distinctively cognitive status of the human brain

Premise (3) Assumption for the sake of argument: LOT* is the appropriate formal architecture fundamental to the distinctively cognitive status of the human brain

Premise (4) SMA: there is a mapping between this bucket of water and LOT*, so the bucket of water is an implementation of the appropriate formal architecture for mentality

But nothing particularly disastrous follows from the above, without further strengthening premise 2 with the additional

Premise (5) CST: *any* physical system implementing the appropriate formal architecture for mentality *is thereby* a mental system.

Conclusion Therefore, this bucket of water *is* a mental system, in **contradiction** with premise 1.

Hence what SMA directly threatens, and what has served as the implicit fulcrum in the trivialization controversy, is not simply a technical paradigm wherein computation is deemed to provide the appropriate mathematical framework for a scientific study of the mind, just as, say, Hilbert Space is the appropriate mathematical framework for quantum mechanics, and 4-dimensional Minkowski geometry for special relativity. Instead, what is threatened is a very specific version of CTM which is committed to CST. As above, defenders of CTM typically try to block the *reductio* argument by rejecting SMA along with premise (4). In contrast, I would advocate rejecting (5), which is a theoretically objectionable tenet in any case. CST is overly simplified and far too strong, and I will argue that the generic rendition of CTM in premise (2) should instead be elucidated in a much more expansive and empirically plausible manner.

3 CST Rejected

It is worth noting that from a normal scientific perspective, CST appears curious indeed. There are many different levels of description and explanation in the natural world, from quarks all the way to quasars. But there is no comparable 'sufficiency thesis' in chemistry, biology, geology, astronomy.[1] In other special sciences, inclusion in categories at the relevant level of description is a matter of degree and scientific utility, and often requires conventional choices for applying taxonomical rubrics in borderline cases—it's not a matter of some tidy and uniformly applicable sufficient condition or 'intrinsic' property. For example, 'being a tectonic plate' is not

[1] An anonymous reviewer has observed that this claim is reminiscent of remarks made in Searle (1980) against Strong AI, wherein the latter position is committed to CST. And while there is definitely some common ground here, in that I agree with Searle in rejecting CST, the underlying reasons and positive stand being endorsed are quite different. Searle has an extremely robust view of Original Intentionality as an essential property of minds, and as indicated in an ensuing paragraph, I disagree with Searle on this point, and defend a much weaker and 'operational' view of the mind in terms of an evolved ensemble of cognitive capacities. Searle invokes causal powers because he holds that Original Intentionality is a type of unique phenomenon ultimately produced by the causal powers of the brain. In contrast, I do not endorse such a 'reified' view of mentality, and only hold that salient causal powers are required in order to manifest the relevant *capacities*, rather than to reproduce some elusive organic bi-product.

considered to be an intrinsic property of the conglomeration of particles categorized as such, and this geophysical level of description is not captured by any simple sufficiency thesis.

Similarly there is no ‘cerebral sufficiency condition’ for when an accumulation of cells, neurotransmitters, etc. comprise a brain. This is a biologically expedient category, but it’s not an intrinsic property of material configurations. However, even though there is no simple and tidy sufficiency condition for ‘being a brain’, CST nonetheless makes the rather remarkable claim to provide a hard and fast sufficiency condition for when such a semi-determinate lump of matter constitutes a *mind*. And I would diagnose much of the disagreement over CTM, the trivialization arguments, and concomitant defensive critiques of SMA to be engendered by ill advised allegiance to CST. In doing so, the CTM camp places far too great a theoretical (ideological?) burden on computation.

Why? In this regard it’s relevant to observe that both Searle and his CTM opponents *share* a crucial Philosophical commitment regarding what is required of an adequate theory of the mind. Part of what motivates acceptance of CST by advocates of CTM is that such theorists apparently view ‘being a mind’ as an intrinsic property of certain physical configurations such as the human brain. And this drives CST, because it is intended to theoretically capture this feature. Searle also thinks that being a mind is an intrinsic property, and this common ground then supplies a shared platform for disagreement about the role of computation. Searle argues that computation is not intrinsic to physical systems, which he takes to constitute a refutation of CTM, since he holds that the defining mental characteristic of ‘Original Intentionality’ cannot be explained by appeal to some *extrinsic* property or level of description.

And I agree with Searle that computation is not intrinsic to physical systems. However, this is not enough to refute a more plausible, non-CST version of CTM, since I disagree with both Searle and his CST opposition that being a cognitive system is an intrinsic property of various arrangements of mass/energy. This is because, just as in the case of ‘being a brain’, there is no clean and simple sufficiency condition—instead it is a question of *degree* with respect to a number of diverse factors and capacities along many different axes, complexity of organization, etc. Bechtel (1993) aptly characterizes ‘*cognition*’ to include all those processes which occur in organisms that make it possible for them to acquire information from their environment and produce actions in response. There is a branching evolutionary continuum stretching from amoebas (or even slime mould and bacteria?) through molluscs, insects, reptiles, birds, cats, monkeys, dolphins, apes, humans… There is no non-arbitrary point at which to draw the line which cleanly divides mental from non-mental. And the situation is made even more nebulous with the contemporary emergence of various forms of Artificial Intelligence.

Hence to return to the claims of CST, how could the mere fact of implementing the ‘right’ type of *abstract procedure* be enough to magically transform a mindless arrangement of matter into a genuinely mental system? In contrast, I argue that much more is required—the system must be anchored in and interact with the real world in a host of rich and multifaceted ways not satisfied by a mere stone or a bucket of water. There is a wide spectrum of empirical evidence that must be taken into account, and it is ultimately a matter of gradient with respect to a broad ensemble of capacities, rather than just a simple binary yes/no question regarding the satisfaction of an abstract and a priori computational definition.[2] Indeed, the view that mentality is an intrinsic property of certain material configurations would appear to stem from (a covert and residual) commitment to the traditional Cartesian view, quite ill suited to a contemporary naturalistic framework.

4 Within a Given Explanatory Project the Data Set Matters

In terms of a computationally based science of mind, a number of pragmatic and application-specific considerations must come to the fore, to augment the bare and theoretically neutral framework provided by MTC and SMA. As in other branches of empirical inquiry, the starting point should be the set of data we are attempting to explain.[3] In the case of a computational theory of mind, various *cognitive capacities* manifested in terms of intelligent behavior, by, e.g., normal humans. These include things like language acquisition and use, planning a future course of action, control of plan execution, acquiring new non-habitual task behaviors, the alteration of one's actions in non-random correlation with salient properties of objects interacted with in the past, capacity for highly structured behavior such as playing games defined by rules, etc. (e.g. as in Rupert 2018).

One need not subscribe to behavior*ism* in order to acknowledge that behavior nonetheless constitutes a central and indispensable form of evidence. But CTM's strong traditional aversion to behaviorism seems to have made it curiously myopic with respect to the central importance of the phenomena to be explained. On the assumption of CST, CTM is so far removed from the relevant data set that its theoretical stance is critically undermined by SMA and a bucket of water (!). In contrast, one would expect a *genuine* counterexample to a robust version of CTM to be, say, a system that can exhibit the right sort of evidence, but which nonetheless can't be explained through appeal to computational processes (perhaps because its full capabilities violate the Church-Turing thesis and transcend the limits of computability). But rather than being able to address the triviality critique with the obvious rejoinder that the bucket of water is irrelevant, because it's not a cognitive system to begin with and exhibits no capacities in need of explanation, defenders of CST must

[2] The same anonymous reviewer has also rightly noted that there appears to be some alignment here with the central claims of Embodied, Embedded and Ecological cognitive science. And while I am certainly not unsympathetic to such views, in the current context I would prefer to simplify the comparison with the original CTM targets of the triviality critique, and thus adopt a more traditional view. The cognitive system must be physically implemented and able to exchange concretely specified inputs and outputs with the external world, but the respective I/O boundaries demarks the cognitive system per se, and the external factors are salient only to the extent that they can condition the inputs to the system (given its outputs) in important and complex ways. But the core cognitive processing is still centralized and internal.

[3] I articulate this as a rather basic principle of theoretical explanation in general. In Shagrir and Bechtel (2017) a more specific version/application of this general principle is put forward in the particular context of Marr and Computational explanation.

instead argue that it isn't 'really' an *implementation* of LOT*. So the key explanatory burden is placed on an ontologically loaded implementation relation, rather than being driven by salient aspects of the phenomena under investigation. And as will be argued in Sect. 9, this *inverts* the explanatory picture.

As a provisional, rough-and-ready standard for the purposes of theoretical discussion, it might be thought a good indicator of 'genuine mentality' if a system is able to pass the linguistic and robotic Total Turing Test (TTT) (Harnad 1991). Clearly a stone will never meet this condition and will be ruled out from the start. 'Deviant' realizations, meant as counterexamples to CTM, are unable to pass even the minimal standards of the purely conversational Turing test (TT) of (1950), precisely because they lack the ability to manifest evidence in the form of intelligent language use. In order to pass the merely verbal TT, the computational device must produce natural language sentences as output, and Searle's wall can't do this. It may output some electromagnetic radiation that we could further interpret as code for the appropriate sentences, but then we as observers are performing an extra and essential step of interpretation wherein the real cognitive work is done. At least in Searle's hypothetical Chinese Room scenario, the set-up has the ability to interactively process the relevant input patterns and produce output in recognizable/readable Chinese syntax—it *can* pass the purely verbal TT.

Hence when it comes to *scientific explanation* as opposed to merely abstract formal considerations, we will need to place concrete constraints on the specification of the inputs and outputs. The 'symbolic' formal inputs and outputs require canonical interpretations in terms of phenomena in the actual world, such as appropriate *linguistic* output in response to input *questions*. So an initial scientific/pragmatic constraint on the domain of the implementational mapping function $\boldsymbol{M}$ is that the purely *abstract* inputs and outputs of the formalism must be the image of *concretely specified* input stimuli and output behavior of actual cognitive systems. So if we let $\boldsymbol{M}^{-1}(x)$ informally denote the inverse of the mapping function, then $\boldsymbol{M}^{-1}(\mathbf{I}_{\mathrm{F}})=\mathbf{I}_{\mathrm{C}}$, where $\mathbf{I}_{\mathrm{F}}$ is the abstract formal input while $\mathbf{I}_{\mathrm{C}}$ is the actual input stimulus in question, and $\boldsymbol{M}^{-1}$ $(\mathbf{O}_{\mathrm{F}})=\mathbf{O}_{\mathrm{C}}$, where $\mathbf{O}_{\mathrm{F}}$ is the formal output while $\mathbf{O}_{\mathrm{C}}$ is in turn the actual output behavior.[4]

This places very significant restrictions on possible implementations, since systems that do not exhibit the appropriate $\mathbf{I}_{\mathrm{C}}/\mathbf{O}_{\mathrm{C}}$ profiles will be precluded from the start. And this is a natural and seemingly obvious strategy, given that our *reason* for utilizing computation as a mathematical tool is to *explain* a certain category of phenomena. Godfrey-Smith (2009) calls the inclusion of concrete inputs and outputs the specification of a FSA (finite state automaton) 'in the broad sense', and notes that it is the reading most favorable to functionalism and CTM. In a somewhat more prosaic context, he gives the example that if we wish to depict a coin operated Cocacola vending machine in FSA terms, then the inputs will need to be actual *coins*, and the output actual *cans of coke*. And in this context the syntactic expression 'cans of

[4] At some point will potentially need a caveat concerning input/output boundaries relative to 'control system' versus peripherals/transducers. But in principle this should not present a difficulty, once CST is rejected and our explanatory project is instead governed by pragmatic scientific constraints.

coke' *means* real cans of coke, and cannot be given deviant interpretations such as flasks of vodka or thimbles of tea.

However I would further comment that this specification of a FSA 'in the broad sense' isn't just the specification of a FSA *simplicitor*, but rather is the specification of a FSA along with part of its *intended interpretation*, relative to the pragmatic concerns of certain human agents. And it is worth emphasizing that this is a huge departure from a global and interest-neutral approach to computation and formal systems, and constitutes a very significant and context dependent limitation on the principle of Multiple Realizability. The Mathematical Theory of Computation itself places no such restrictions on the interpretation of formal syntax. Accordingly, the inputs and outputs are purely symbolic, and can be interpreted in any *possible* manner consistent with the structure of the formalism. So from the perspective of MTC, this initial move of requiring that the inputs and outputs assume a chosen concrete identity, is comparable to, say, 'solving' the Skolem paradox in model theory simply by *stipulating* that the domain of discourse must be the real numbers.

In contrast to purely formal or mathematical contexts, within the empirical confines of a given explanatory project it *is* justifiable to restrict the range of application to our field of interest, and hence by pragmatic fiat *decree* that we are only concerned with the intended interpretation. However, this doesn't constitute a conceptually illuminating or metaphysically binding constraint on the physical implementation relation per se, and is not a defensible strategy for blocking non-standard interpretations as 'false' or 'non-genuine' in any deep or interesting sense. Such pragmatic restrictions cut no significant Philosophical Ice, and instead merely indicate that these other interpretations are *not relevant* to our particular realm of application.

5 Mesh Between Formal and Causal Structure

The foregoing section has argued that, as a first step, a scientific/explanatory project must invoke concrete, domain specific constraints on the salient inputs and outputs, in order to mesh with the actual data set. This step is a broad spectrum 'behavioristic' calibration with the phenomena under investigation, and will help serve to isolate the type of physical systems that we are attempting to explain. Systems such as tornados and whirlpools do not exhibit any cognitive input/output capacities and hence are filtered out. These systems may be relevant to alternate theoretical endeavors such a classical mechanics or fluid dynamics, but not to cognitive science. In turn, other systems such as humans, chimpanzees and advanced computational artifacts *will* fall within the appropriate behavioral domain.

CTM is generally conceived as a theoretical advance over behaviorism, in that it is concerned not just with black-box stimulus/response patterns, but crucially, with the *means by which* the inputs are transformed to yield the outputs. Internal processing structure is the hallmark of computational accounts, and hence once the systems of interest have been identified in terms of salient behavioral capacities, the requisite next step concerns the pathways *connecting* the inputs and outputs, and hence with the mapping between the computational formalism and the internal state transitions of the physical system. There are any number of different perspectives and levels for describing the very same physical system, and none of them is objectively privileged or fundamental to the system *as such*—instead it remains a question of human choice relative to our interests and goals.

For example, a traditional spring-driven analogue clock can be formally modelled at various *microphysical* levels—at a subatomic level in terms of quantum mechanical processes and interactions, and at a higher microphysical level in terms of molecular thermodynamics. In the latter case, it could also be described in more abstract functional terms as a temperature detector, where the mean molecular kinetic energy of its metallic components tracks the ambient atmospheric temperature. And it can be described and modelled at various *macrophysical* levels as well, such as an intricate classical mechanism with states evolving in accord with continuous real valued equations. It could also be described in more idealized *conventional* terms, where certain selected continuous features are broken into discrete segments and given a chronological interpretation. And yet again, this relatively advanced design level stance could be ignored, and the object could be given a more rudimentary functional depiction, e.g. where its size and inertial properties make it useful as a doorstop.

If we want to model a coin operated coke machine using an FSA, then we are only concerned with the concrete device *described as* a properly engineered and functioning coke dispenser, and not in any of the many other possible alternative descriptions of the very same physical system. Hence in addition to the initial 'behavioristic' specification that the inputs are actual coins (in some national currency) and the outputs real cans of coke, it is also requisite that the internal causal pathway leading from input coin to output coke is respected by the implemenational mapping function $\boldsymbol{M}$. In other words, the individuated physical states comprising the domain of $\boldsymbol{M}$ must correspond to the relevant engineered states in the causal chain leading from input to output, and the mechanism's internal physical pathway reflected in the matching state transitions of the computational procedure as specified by $\boldsymbol{M}$.

Similarly, if interpreting the brain as a computational device is to have scientific utility, then the ascribed formal structure must correspond to aspects of the relevant causal structure i.e. the causal structure that enables it to behave in ways salient to its status as a *cognitive* system, as opposed to a coke machine. So (very generally and not at this point worrying about the many intervening levels of description connecting bottom-up machinery with top-down cognitively characterized behavior) the internal computational pathway mediating *formal* inputs $\mathbf{I}_{\mathrm{F}}$ and outputs $\mathbf{O}_{\mathrm{F}}$ must integrate with the physical causal pathway from *concrete* inputs $\mathbf{I}_{\mathrm{C}}$ to outputs $\mathbf{O}_{\mathrm{C}}$. In order to give an explanatory account of the mind-brain, utilizing high level computational description, we need to treat the brain along the lines of a biologically engineered device comparable to one of our computational artefacts, since many of the same kinds of pragmatic constraints will need to be invoked.

6 CTM *sans* CST

As an illustration, we will again take LOT as a paradigmatic instance of CTM, and hence as a prime target for the Putnam-Searle triviality critique. According to classical LOT, mental processes are explicitly viewed as formal operations on a linguistically structured system of internal symbols. Additionally, the LOT incorporates the belief-desire framework of psychological explanation, which holds that an agent's rational actions are both *caused* and explained by intentional states such as beliefs and desires. On the LOT model, these states are sustained via sentences in the head that are formally manipulated by the cognitive processes which lead to actions. Hence propositional attitude states are treated as computational relations to sentences in an internal processing language, and where the LOT sentence serves to represent or encode the propositional content of the intentional state.

And merely for the sake of argument, let us temporarily suppose, as before, that cognitive scientists eventually endorse some suitable version of LOT as the underlying functional/computational architecture of the human mind, and thus hold that the brain serves as a biological implementation of this formal structure. The proposed 'thought experiment' is not an attempt to foretell the outcome of future findings, nor to make empirically loaded speculations from the comfort of an arm chair. Indeed, my overall goal is not to argue that CTM (of any variety) is *in fact* the right approach—this remains an open question which must be settled by future scientific research (and even assuming CTM, it may well turn out that the traditional belief-desire framework enshrined in LOT will be entirely abandoned by successful theories, as many have contended). Instead, its conceptual job is merely to provide a counterexample to the triviality arguments. So, as long as the envisioned state of affairs is a conceptually cogent scenario, it shows that an SMA account of physical implementation is consistent with the possibility of empirically rich and scientifically fruitful versions of CTM, and hence that the broad-sweeping trivialization strategy is logically *invalid*. For the sake of illustration I use LOT as an exemplar of the classical approach, although the result applies to computational architectures in general.

The typical human being is able to exhibit intelligent linguistic behavior. Relative to the particular explanatory project of cognitive science, our aim is now to explain this capacity in terms of the causal powers of the brain when viewed as an organic hardware device running LOT as its cognitive software. Thus we are interested in the multitude of physical processes taking place inside the skull at the level of physical organization *as a brain*, and further we are interested in an even higher level of description of the brain as performing computations that account for its cognitively salient capacities. Obviously, there would be no scientific value in mere ad hoc mappings between LOT and the bare lump of mass/energy occupying the cranial cavity. As in the case of computationally modelling a coke machine (and unlike trivialization exercises), a scientifically significant approach is not free to view this complex physical system in terms of brain-irrelevant aspects like cosmic ray bombardment, gravitational fields, electron spin, arbitrary molecular kinetics, etc. Although in principle many such mappings will be simultaneously possible, *a la* SMA, a theoretically fruitful and substantive approach must restrict itself to salient *causal* factors pertaining to the physical system's time-evolution, when viewed as a brain able to produce the type of linguistic input/output patterns under study.

So there will be a myriad of pre-existing and empirically intransigent 'wet-ware' constraints that the mapping will have to satisfy, in order to respect the salient causal structure of brain activity as discerned by neuroscience. The largely independent body of functional and anatomical data would then supply a host of highly non-trivial restrictions on how the physical system itself is characterized and what the material state transitions should look like, that are interpreted as implementations, under $\boldsymbol{M}^{-1}(x)$, of the abstract computational procedures. Such structural constraints may resemble the requirements of a 'mechanistic' view of computation, as proposed, e.g., by Piccinini (2015a, b), Dewhurst (2018), Mollo (2018), but a vital difference is that the current constraints are put forward only *relative* to a given explanatory project that employs computation as a mathematical tool, and not as having anything whatever to do with a general theory of 'real' computation. Thus the principle difference is that these structural/mechanistic aspects are not motivated by an attempt to provide necessary or sufficient conditions for 'genuine' as opposed to 'spurious' implementations, but rather are driven merely by the explanatory requirements for a plausible *scientific account* of the brain (as in Schweizer 2012). So the constraints are fundamentally brain-structural rather than computational, and I do not posit the need for mechanistic 'digits' and 'processors', nor any other rigid or preconceived computational restrictions on how the physical system should be depicted. Instead, the envisioned constraints are driven purely by a scientific analysis of the *brain* as a physical mechanism.

If a physical system when viewed as a brain were methodically interpretable as implementing the LOT, this would entail that the transitions between the various neurological states instantiating respective tokens of mentalese symbols obeyed a causal progression in accord with the transformation of these symbols as prescribed by the abstract computational formalism. Thus, in rather simplistic terms, if the currently implemented state of the LOT formalism entails the transition to abstract computational state $\boldsymbol{A}_i$, and if $\boldsymbol{M}^{-1}(\boldsymbol{A}_i)$ is the concrete neurological state $\boldsymbol{C}_j$, then a testable consequence of the theory is that the underlying causal structure of the brain will next produce state $\boldsymbol{C}_j$. And *if* this were the case, it would provide a scientifically fruitful and explanatorily powerful key to organic cognition, because it would constitute a unifying perspective tying together actual brain function and the standard belief-desire framework of intentional explanation.[5]

This abstract computational interpretation of brain activity would also need to mesh with the input and output capabilities that we want to explain via the attribution of internal cognitive structure, e.g. intelligent linguistic performance, as in a Turing test. So the computational level of description would have to conform with observable input and output patterns, such as sentences in an English (or Chinese) conversation, to yield successful predictions of both new outputs given novel inputs, *and* predictions correctly describing new *brain configurations* entailed by the theory as realizations of the appropriate formal transformations required to produce the

[5] In the case of an artifact, such as a coke machine, this type of enterprise mapping causal to computational structure is obviously much more straightforward, since the physical device was intentionally designed with a particular formal blueprint already in mind. In the case of a biologically 'engineered' system such as the human brain, this 'canonical' top-down perspective is not applicable. Hence for this type of (still quite speculative) CTM project to be successful, there would need to be a good deal of reflective equilibrium and interplay in depicting both causal and computational levels of description to achieve a mutually compatible mesh.

predicted output behavior. Hence it would have to yield confirmable and fully integrated predictions on *two* distinct levels of analysis.

As above, there is no 'cerebral sufficiency condition' for when an accumulation of cells, neurotransmitters, etc. comprise a brain. This is a biologically expedient category, but it's not an intrinsic property of material configurations. And similarly, in the present thought experiment, the brain serves as a biological implementation of the formal LOT structure, but still, this is *not* an intrinsic property of the brain as a complex biochemical device. It simply means that this particular LOT mapping is successful at underwriting predictions of future events at the salient level of description. But the brain can be interpreted as simultaneously implementing any number of different computational formalisms, and none is intrinsic. Furthermore, there is no reason to believe that the interpretation of the brain as implementing LOT constitutes a *unique* mathematical solution to the constraint satisfaction problem posed by our explanatory aims. Indeed, scientific theorizing is an activity carried out by fallible human theorists, working with severely restricted data, and deploying their own limited cognitive resources. Abstract scientific theories are accurate only as a matter of degree—they are defeasible approximations corrected and improved over time.

Nonetheless, if the foregoing project were to a large degree accomplished, it would have exceedingly non-trivial scientific/empirical value. And this value is not in the least undermined by Putnam-Searle type mappings—objections of this kind have polemical force only on the assumption of CST. But in light of the many *concrete empirical* constraints and opportunities for testing predictions of both external behaviour and internal brain state, the CST is rendered a completely gratuitous consideration. There is no single and simple sufficiency condition in this highly complex and multifaceted scientific enterprise, and merely implementing the LOT does not magically convert matter into mind. On the more scientifically plausible version of CTM currently envisioned, computation supplies a successful high level mathematical description of the brain for the prediction and explanation of actual events, including the production of utterances in some natural language conversation. In contrast, a tepid bucket of water manifests no actual events in need of cognitive explanation. Thus if faced with the triviality objection that there is nonetheless a level of description at which the bucket of water can be interpreted as an implementation of LOT, the advocate of CTM *sans* CST can happily shrug and respond—'Yes, and so what?'

7 The Computational Stance

In this manner we adopt what could be termed a 'Computational Stance' towards physical systems (as in Schweizer 2019). This approach is in central ways comparable to Dennett's (1981) Intentional Stance, wherein intentional states such as beliefs and desires are not posited as objectively real phenomena, but instead are treated as mere 'calculational devices' or '*abstracta*' in Reichenbach's sense (like point masses and perfectly frictionless surfaces in classical mechanics), used to predict observable events, but without any additional ontological commitments. Analogously, I would construe abstract computational states on a similar footing. In the case of our purpose-built artifacts, these abstract states are *idealized* formal notions that we employ to describe such devices from a higher design-level perspective. Classic digital computation is rule-governed syntax manipulation, and as such is no more intrinsic to physical configurations than is syntax itself. Furthermore, discrete states are idealizations, since the physical processes that we interpret as performing digital computations are continuous (in the standard non-quantum case). Thus discrete states do not literally correspond to the underlying causal substrate. We must *abstract away* from the continuity of actual physical processes and impose a scheme of conventional demarcations to attain values that we can then *interpret* as discrete. Hence this elemental building block of digital procedures must be projected onto the natural order from the very beginning (as Turing observed in 1950), and in this respect is a convenient fiction rather than a literal depiction.

Dennett holds that there is no internal matter of fact distinguishing systems that 'really' possess intentional states from those which do not—the strategy only requires us to view the system *as if* it possessed such states. Hence there is nothing in principle to stop one from depicting a stone as an intentional system if one so chooses. In a similar vein, I would argue that there is no deep or metaphysically grounded fact regarding whether or not a physical system 'really' implements a given computational formalism. In the case of artifacts such as my desk top computer, I can gain a huge increase in the ability to predict (and exploit) its future states if I adopt a computational stance as opposed to viewing it as a brute physical mechanism. And this is because it has been designed and constructed for exactly this purpose. In contrast, a stone has not been so designed, and the pragmatic value of viewing it in computational terms will be rather limited.

Relative to particular goals and design parameters imposed by human engineers, in conjunction with known principles of materials science, there can be very tightly constrained abstract solutions at particular levels of description, e.g. circuit theory (Scheutz 1999). SMA does not imply that such mappings are 'arbitrary', and surely the impressive success and reliability of our artifacts is not a subjective phenomenon. As with Dennett's Intentional Stance, predictive success is an objective criterion. However, to the extent that success *is* achieved in the case of our artifacts, it ultimately rests upon skilled manipulation of the physical substrate. And the ever present possibility of error and malfunction indicates that an abstract computational description of this (continuous) substrate is still a normative idealization and not an 'intrinsic' characterization.

In a related and potentially compatible vein, Millhouse (2019) proposes that the Dennettian notion of 'real patterns' be embraced as the theoretical criterion for physical computation. And I would agree that such an approach can provide a very useful, yet metaphysically modest, handle for comparing the degree to which various computational depictions of physical systems possess potential pragmatic/epistemic value. Millhouse argues for this approach along the standard motivating lines—as a defense against trivialization and the concomitant threat to CTM. In contrast, I would advocate relinquishing tacit commitment to CST and then no defense is required. The algorithmic simplicity criterion advocated by Millhouse has the virtue of being able to provide an objective formal comparison between computational interpretations of physical systems, without needing to invoke the metaphysically dubious (and I would argue, ultimately unsuccessful) standards of the semantic or mechanistic approaches. But still, I would see this as a fundamentally pragmatic rather than ‘intrinsic’ or ‘realist’ criterion. For a limited span of time and to some partial degree of accuracy, the abstract computational description can be used to track the underlying causal dynamics of the physical system. Thus (as in Schweizer 2012) physical implementation is not a binary yes/no question, but rather is a matter of degree.

And when it comes to applying our abstract mathematical formalisms to the physical world, this sort of criterion is obviously of immense practical value. As noted at the outset, electrical engineering is not ‘threatened’ by SMA. And no one thinks that interpreting a stone as implementing various formal procedures will provide us with an epistemic pay-off, and neither will interpreting a bucket of water as implementing the LOT provide us with a good explanation of anything. In contrast, characterizing a complex physical artifact at the level of circuit theory and applying abstract computational ascriptions can be of enormous practical benefit. But there is nothing physically or metaphysically privileged about circuit theory as a level of description, and it does not preclude alternative characterizations and different computational mappings ascribed to the very same physical system. Hence such pragmatically ‘favored’ mappings are in no way inconsistent with SMA and the basic Computational Stance.[6]

8 Is Computational Implementation a Necessary Condition?

Searle has forcefully maintained that implementing a computational formalism is not a *sufficient* condition for mentality, and I would agree. The paper has argued that much more is required, and that the computational sufficiency thesis, which serves as the implicit fulcrum of the triviality debate, should be abandoned by the CTM camp. But it may still seem a theoretically significant question, in this broader CTM context, to ask whether or not computation nevertheless supplies a *necessary* condition? This would appear to still provide serious theoretical traction for a computational approach to the mind.

However, even Searle can concede that computation *is* a necessary condition, but only vacuously so. Given the liberality of SMA and his attendant trivialization critique, any mental system can be interpreted as implementing virtually any computational formalism. So being an implementation is necessary in the rather mundane sense of being unavoidable. This line of reasoning also underwrites Putnam’s criticism that if we specify concrete inputs and outputs for the cognitive system, as in Sect. 4 above, then CTM collapses into a covert form of behaviorism. This is because we can map any computational procedure we like to the intervening

[6] Lee (2018) defends an interesting view of computational ’pluralism’ that also allows multiple computational interpretations. However, the version of ’anti-realism’ that I advocate is purely MTC/SMA based, and thus fundamentally concerned with syntactic mappings, whereas Lee’s pluralism is with respect to mechanistic versus semantic accounts, neither of which I would endorse.

physical processes connecting input and output, and thus the attribution of internal processing structure adds nothing significant to our explanation. But this line of critique is itself trivialized by the current observation that SMA is plainly too weak a standard for our purposes.

In the confines of a scientifically viable theory of the mind which invokes computation as its signature mathematical framework, the global and theoretically neutral SMA must be augmented by additional factors and domain specific considerations, as in the preceding section. Let us label this type of pragmatic and theoretically specialized approach an Explanatory/Causal Mapping (ECM). It does not purport to constitute a global, 'realist' account of the physical implementation relation in general, but instead supplies a *pragmatically filtered subspace* of implementations that is relevant to our particular CTM project. And thus proffered counterexamples to CTM such as stones and buckets of water have been filtered out. An obvious question to ask at this point is—does *ECM* then supply the basis for a necessary condition for mentality? Does this more empirically plausible, non-CST version of CTM entail that any genuinely mental system must fall within the bounds of the pragmatically filtered subspace? If so, then perhaps CTM would possess the virtue of being able to identify a type of 'real pattern' via the attribution of computational structure.

Hence this would seem to constitute a potentially promising line of inquiry. But alas, it turns out that many (if not most) purported cases of computation in naturally occurring systems (as opposed to our custom-designed artifacts) fail to meet even the minimal requirements of a simple mapping account. This is because they lack a detailed mapping function to begin with. In particular, they lack a specification of intervening computational state transitions between input and output. There *is* a mapping of inputs and outputs, and there is an intervening *physical* process connecting them, so that the physical process is in effect treated as 'computing' a black-box function-in-extension. But there is no algorithmic or abstract procedural detail, and hence the *formal method* by which the function is computed is left entirely unspecified. Of course, *a la* Searle and SMA, we *could* map the intervening physical process to any number of different algorithmic pathways, and hence provide the details of particular implementations that *could* be said to compute the behavioristic function in extension. However, this would serve no purpose in a scientific attempt to *explain* cognitive behavior, and it does not satisfy the requirements of ECM. Unbridled SMA is typically dismissed by those in the CTM camp as being in principle too weak a criterion, but quite ironically and perhaps surprisingly, in most cases the favored computational account does not even meet the basic SMA requirements.

As Rescorla (2017) rightly observes, in stark contrast with our computational artifacts, many cases of purported computation in the *natural* world, even in the specific context of cognitive science, conspicuously lack *any* intermediate syntactic, algorithmic or formal level of description. For a salient case in point exposited by Shagrir (2014) (and not canvassed by Rescorla), consider the neural integrator in the oculomotor system. The scientific account given is that the system produces eye-position codes by *computing* mathematical integration over eye-velocity encoded inputs, thereby enabling the brain to move the eyes to the correct position. It seems clear that the extensional pairs of eye-velocity inputs converted to eye-position-outputs in the case of the ocular mechanism can be captured *intensionally* with the mathematical function for integration. But which algorithm does the brain use to *compute* the integration function, and how are the abstract state transitions entailed by the formal procedure mapped to intervening brain states? Unless these details are provided, the sense in which computation is involved remains unclear. In the case of the visual system, there is compelling reason to view the internal brain process as mirroring or calibrating itself with distal factors in order to successfully control eye position. But from this alone it does not follow that the explanatory description is explicitly *computational*.[7]

Rescorla (2017) is overtly sanguine about this general state of affairs. He identifies what he calls three levels of 'computational' description: the representational, the syntactic and the physical/mechanical, and as above, rightly notes that in the case of cognitive systems, the intermediary level of syntax is generally missing. He provides the example of mammalian cognitive maps (in rats) as illustration, where scientists describe the maps in high level representational terms, and explore rats' neural mechanisms serving as implementations, with no intervening syntactic or algorithmic characterization. But again this raises the question—in exactly what sense does the physical processes in question implement a well-defined *computational* procedure? Or is this just 'computation' in an equivocal or purely *metaphorical* sense?[8]

I would certainly agree that there are three generally recognized levels of description, but would argue, in accord with the fundamental MTC perspective (along with, e.g. Chalmers 2012; Millhouse 2019) that it is a syntactically specified *effective procedure* which constitutes the distinctively *computational* level. In contrast, the physical/mechanical level concerns the particular *implementation* of the effective procedure in question, while the 'representational' level concerns the intended *interpretation* of this procedure. Thus it is the syntactical, algorithmic level which is fundamental, and which defines computation as such. So for a properly computational description of some system, it is necessary to specify the *internal state transitions* which constitute the algorithmic pathways from input to output, and which individuate the particular procedure being ascribed. In this respect, the Computational Stance differs from Dennett's original Intentional Stance, in that it should *not* take a macroscopic, black-box perspective on the system—instead, the intervening abstract state transitions are integral to the stance, insofar as they track corresponding *physical* pathways as specified by the mapping function $\boldsymbol{M}(x)$.

[7] Perhaps it will be said that the brain does not implement a classic digital formalism to compute the integration function, but rather is performing an *analogue computation* instead. This is not an implausible claim, but we would still need to be supplied with the specific details, the actual analogue *method* by which the function is being computed in the brain. For example, the differential analyser is an analogue computer, and it works in accordance with well defined principles (Shannon 1941). Similarly in the case of purported analogue computations performed by the brain, we would still need to know the specific details.

[8] Some are inclined to equate 'computation' with a (usually much vaguer) notion of 'information processing', and so might contend that these are cases of computation in the latter sense. However, in agreement with Piccinini and Scarantino (2010), I would respond that the two notions are not synonymous.

Classical computation is formal, rule governed syntax manipulation, and a *definitional* constraint supplied by MTC is that the rules can be followed without any additional interpretation or understanding. Hence effective procedures can be executed without knowing what the 'symbols' in question are supposed to mean. Semantics and 'representational content' are add-ons, and no single or privileged interpretation is determined by a particular effective procedure. As a well known example, consider a Turing machine intended to compute the values of a particular truth function, say inclusive disjunction. The machine itself is a program for manipulating the symbols '0' and '1' on given input tapes, where '0' is intended to denote False and'1' denotes True. As such, it can easily be *reinterpreted* as computing the truth function associated with conjunction instead of disjunction, simply by flipping the intended reference of the manipulated symbols so that '0' denotes True and '1' denotes False. There is no independent fact of the matter regarding what these syntactic tokens 'really mean'—their referential value is dependent upon a scheme of interpretation which is not itself specified or determined by the computational activities of the Turing machine. The formal behaviour of the device is the same in either case, and the rule governed procedure can be executed with no projected interpretation at all.[9]

In addition to being multiply interpretable, computational formalisms, as above, are multiply realizable. In the purely global and theory neutral case, SMA places no special restrictions on the physical domain of the mapping function. However, within the context of a computational theory of mind, ECM has been identified as the appropriate filter. And this added requirement avoids Putnam's trivial 'collapse into behaviorism' critique of CTM, which only works through appeal to SMA. If the story must invoke a robust correspondence with the actual causal pathways leading from concrete inputs to concrete outputs, then the stance is not behavioristic, but instead is explanatory/causal, and where *computation* provides the integrating, high-level mathematical key. The ECM approach thus avoids behaviorism, and it also avoids the spectre of neuro-chauvinism, since there is no stipulation that physically alternative systems could not serve as implementations of the same computational blue-print. The behavioral data set must remain somewhat regimented, to the extent of manifesting recognizable cognitive capacities, but ECM still allows Multiple Realizability in terms of the causal pathways that *could* serve as implementations of the intervening abstract procedural pathways. ECM thus steers a plausible course between the two poles of behaviorism and neuro-chauvinism, which Godfrey-Smith (2009) notes has been one of the chief theoretical challenges for a computational approach to mentality.

9 Locus of Explanatory Force

It's important to reiterate that the ECM invocation of causal structure is not imposed as a necessary or global constraint on physical implementation per se, and should not be conflated with a tacit appeal to general causal, functional or mechanistic

[9] These and related global criticisms of the semantic view of computation are put forward in Schweizer (2017).

accounts of computation in physical systems. On my approach, correspondence with causal/functional mechanisms is invoked only with respect to an *explanatory theory of* a particular domain which adopts computation as a formal tool and utilizes testable predictions to establish its scientific credence. And this move can be justified only by its instrumental utility—it makes no dubious metaphysical claims about 'real' implementation or the 'intrinsic' purpose or 'proper function' of causal mechanisms. Instead such factors are motivated simply by generic aspects of scientific explanation (e.g. ability to support testable predictions) and are not philosophical stipulations regarding the conditions necessary for supposed 'genuine computation' in the physical world. Hence, in my view, Piccinini and other mechanists extract these features from what is required of a good explanation, and inappropriately incorporate them into a global and literal view of physical implementation. In contrast, I argue that the explanatory burden should be shifted away from an ontologically loaded implementation relation (and CST), and instead driven by pragmatic concerns stemming from the particular explanatory project at hand.[10]

Like other branches of mathematics, such as set theory, topology and differential calculus, MTC provides a precise, well defined formal *tool* which can be applied to the physical world in a virtually limitless variety of ways. And just like, e.g., 'being a member of a set', or 'being the output value of an enumeration function on the positive integers', so too 'being an implementation of a computational formalism' is not a metaphysically grounded or intrinsic property of physical entities, but rather is a highly abstract level of description that we project onto the world according to our interests and diverse pragmatic goals. It is founded on an observer-dependent act of ascription, upon an entirely *conventional* correlation between physical structure and abstract formalism. Furthermore, this conventional mapping is essentially prescriptive in nature, and hence projects an outside *normative* standard onto the activities of a purely physical device. And I would advocate this type of 'anti-realist' pragmatic perspective, in lieu of attempts to give overarching semantic, causal or mechanistic constraints purporting to distinguish literally 'true' from 'false' cases.

Pragmatic factors do not need or claim to support global and uniform necessary conditions for implementation (and the ever present non-zero probability of error in physical systems indicates that there is no fully *sufficient* condition, either). Different operational desiderata will have shifting roles and prominence in different contexts of application, and will be satisfied to varying *degrees* dependent on the goals and purposes in question, as well as the state of our technological progress. Computation is a highly versatile tool, and there is no single and objective class of phenomena that can be isolated as comprising the 'real' instances of physical implementation. Instead, SMA specifies the maximal and context neutral space of possibilities, and varying pragmatic considerations can then be applied to carve out different subsets within this space which prove useful or interesting according to our divergent human purposes. In short, physical computation is not a natural kind—it is dependent upon human convention, interpretation and choice.

[10] See Schweizer (2019) for an extended critique of causal and mechanistic accounts.

Thus for a computational *theory of* the mind which rejects CST there will be a 'conservation of explanatory force' entailed, in order for the theory to do its scientific job. But a central issue concerns the *locus* of this explanatory force. One of my basic points is that the various causal and mechanistic factors that may be required by a scientific theory of some given subject area should be *localized* within that particular explanatory project, rather than promulgated as 'necessary' constraints used to fetter the in-principle standards of physical implementation. Contrary to the spirit of CST, the general concept of physical implementation should not bear the explanatory burden—instead the subspace of possibilities relevant to our particular endeavor must be isolated by invoking appropriate constraints on SMA. In the case of CTM an Explanatory/Causal mapping has been specified to define a pragmatically filtered subspace. And in contrast to current terminological practice, for the result to be a clear and unambiguously *computational* theory of the mind, it is perhaps a compelling desideratum that ECM provides a necessary condition, and hence that the salient algorithmic details be explicitly delineated.

10 Conclusion

The paper has shown that the general form of the SMA-based triviality critique is invalid. This move requires the rejection of the computational sufficiency thesis, which I've argued is theoretically unpalatable for a variety of reasons. In place of CST we should adopt a more empirically plausible ECM view on the relation between computation and mentality. On this amended version of the CTM paradigm, it may well turn out to be scientifically profitable to depict the workings of actual cognitive systems in computational terms. However, this does not warrant the a priori CST stipulation that cognitive systems should be computationally *defined*. Computation per se has no mystical powers of 'cognitive transformation', and hence the burden of explanatory force should be shifted away from the relation of physical implementation, and instead be driven by salient aspects of the phenomenon under investigation, as well as by the general requirements of scientific explanation. This position is consistent with a maximally liberal and unfettered view of computation as an abstract mathematical tool like any other, and at the same time is immune to the triviality arguments that invoke the unconstrained and theoretically neutral standards of SMA.

Along the way, I've also supplied some independent reasons for retaining the simple mapping account of physical implementation, and given various criticisms of alternative views. However, whether or not one chooses to accept or reject SMA, the conclusion nonetheless remains that 'saving' (a scientifically plausible version of) CTM can no longer be seen as a compelling reason for rejecting SMA, and for promulgating more restrictive and less mathematically general views on physical implementation. And in any case, it is not to be simply *assumed* that a CTM approach, of any variety, will prove to be the successful paradigm in the science of human mentality—the definitive nature of the explanatory relation between formal computation and human intelligence remains an open question.

What then are the ramifications of this view regarding the closely associated field of *Artificial* Intelligence? Traditional Strong AI is committed to CST, and hence my arguments serve to undermine this type of position for artifacts, just as for humans. However, the traditional, intrinsic view of the mind, held by both Searle and proponents of CST, has, in the current discussion, been replaced by a more operational analysis, wherein the phenomenon to be addressed is seen as a complex ensemble of *cognitive capacities*. Hence if these capacities could be sustained computationally, by a project of Artificial General Intelligence, then such a project should potentially be deemed successful, independently of the theories and methods that might ultimately account for such capacities in humans.[11]

Acknowledgements This paper was originally presented at the IACAP 2018 conference, and I would like to thank Ron Chrisley, Joe Dewhurst, Oron Shagrir, and other members of the audience for useful questions and comments. I would also like to thank two anonymous reviewers for helpful and constructive feedback, from which the final version of the paper has greatly benefited.

References

Bechtel, W. (1993). The case for connectionism. *Philosophical Studies, 71*(2), 119–154.

Bishop, J. M. (2009). Why computers can't feel pain. *Minds and Machines, 19,* 507–516.

Block, N. (1978). Troubles with functionalism. In C. W. Savage (Ed.), *Perception and cognition*. Minneapolis: University of Minnesota Press.

Chalmers, D. J. (1996). Does a rock implement every finite-state automaton? *Synthese, 108,* 309–333.

Chalmers, D. J. (2012). A computational foundation for the study of cognition. *Journal of Cognitive Science, 12*(4), 323–357.

Chrisley, R. L. (1994). Why everything doesn't realize every computation. *Minds and Machines, 4,* 403–420.

Copeland, J. (1996). What is computation? *Synthese, 108,* 335–359.

Dennett, D. (1981). True believers: The intentional strategy and why it works. In A. F. Heath (Ed.), *Scientific explanation: Papers based on Herbert Spencer lectures given in the University of Oxford*. Oxford: University Press.

Dewhurst, J. (2018). Computing mechanisms without proper functions. *Minds and Machines, 28,* 569–588.

Fodor, J. (1975). *The language of thought*. Cambridge: Harvard University Press.

Fodor, J. (1981). The mind-body problem'. *Scientific American, 24,* 114.

Godfrey-Smith, P. (2009). Triviality arguments against functionalism. *Philosophical Studies, 145*(2), 273–295.

Harnad, S. (1991). Other bodies, other minds: A machine incarnation of an old philosophical problem. *Minds and Machines, 1,* 43–54.

Johnson-Laird, P. (1988). *The computer and the mind*. Cambridge: Harvard University Press.

Kripke, S. (1982). *Wittgenstein on rules and private language*. Cambridge: Harvard University Press.

[11] In the present discussion I do not canvass issues regarding the (currently seemingly intractable) phenomenon of conscious qualitative experience, and hence such a stand may not generalize. As a final note, Ron Chrisley has brought to my attention the fact that maintaining a scientifically viable version of CTM without accepting CST is compatible with some of the remarks he makes at the end of his 1994 paper.

Lee, J. (2018). Mechanisms, wide functions, and content: Towards a computational pluralism. *British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy061.

Lycan, W. (1981). Form, function, and feel. *Journal of Philosophy, 78*(1), 24–50.

Maudlin, T. (1989). Computation and consciousness. *Journal of Philosophy, 86*(8), 407–432.

Milkowski, M. (2013). *Explaining the computational mind*. Cambridge: MIT Press.

Millhouse, T. (2019). A simplicity criterion for physical computation. *British Journal for the Philosophy of Science, 70*(1), 153–178.

Mollo, D. (2018). Functional individuation, mechanistic implementation: The proper way of seeing the mechanistic view of concrete computation. *Synthese, 195,* 3477–3497.

Newell, A., & Simon, H. (1976). Computer science as empirical enquiry: Symbols and search. *Communications of the ACM, 19*(3), 113–126.

Newman, M. (1928). Mr. Russell's "causal theory of perception". *Mind, 37,* 137–148.

Piccinini, G. (2015a). Computation in physical systems. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/fall2015/entries/computation-physicalsystems/.

Piccinini, G. (2015b). *Physical computation*. Oxford: Oxford University Press.

Piccinini, G., & Scarantino, A. (2010). Computation vs. information processing: Why their difference matters to cognitive science. *Studies in History and Philosophy of Science, 41*(3), 237–246.

Pinker, S. (1997). *How the mind works*. New York City: W.W. Norton.

Putnam, H. (1967). The nature of mental states. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 1–223). Pittsburgh: Pittsburgh University Press.

Putnam, H. (1988). *Representation and reality*. Cambridge: MIT Press.

Pylyshyn, Z. (1984). *Computation and cognition: Toward a foundation for cognitive science*. Cambridge: MIT Press.

Rescorla, M. (2014). A theory of computational implementation. *Synthese, 191,* 1277–1307.

Rescorla, M. (2017). Levels of computational explanation. In T. Powers (Ed.), *Philosophy and computing: Essays in epistemology, philosophy of mind, logic, and ethics. Philosophical studies series* (Vol. 128, pp. 65–84). Berlin: Springer.

Rupert, R. (2018). Representation and mental representation. *Philosophical Explorations, 21*(2), 204–225.

Scheutz, M. (1999). When physical systems realize functions. *Minds and Machines, 9*(2), 161–196.

Schweizer, P. (2012). Physical instantiation and the propositional attitudes. *Cognitive Computation, 4*(3), 226–235.

Schweizer, P. (2017). Cognitive computation *sans* representation. In T. Powers (Ed.), *Philosophy and computing: Essays in epistemology, philosophy of mind, logic, and ethics. Philosophical studies series* (Vol. 128, pp. 65–84). Berlin: Springer.

Schweizer, P. (2019). Computation in physical systems: A normative mapping account. In D. Berkich & M. V. d'Alfonso (Eds.), *On the cognitive, ethical, and scientific dimensions of artificial intelligence—Themes from IACAP 2016. Philosophical studies series* (pp. 27–47). Berlin: Springer.

Searle, J. (1980). Minds, brains and programs. *Behavioral and Brain Sciences, 3,* 417–424.

Searle, J. (1990). Is the brain a digital computer? *Proceedings of the American Philosophical Association, 64,* 21–37.

Shagrir, O. (2001). Content, computation and externalism. *Mind, 110*(438), 369–400.

Shagrir, O. (2014). The brain as a model of the world. In *Proceedings of the 50th anniversary convention of the AISB, symposium on computing and philosophy*.

Shagrir, O., & Bechtel, W. (2017). Marr's computational level and delineating phenomena. In D. M. Kaplan (Ed.), *Explanation and integration in mind and brain science* (pp. 190–214). New York, NY: Oxford University Press.

Shannon, C. E. (1941). Mathematical theory of the differential analyzer. *Journal of Mathematics and Physics, 20,* 337–354.

Sprevak, M. (2010). Computation, individuation, and the received view on representations. *Studies in History and Philosophy of Science, 41,* 260–270.

Stich, S. (1983). *From folk psychology to cognitive science*. Cambridge: MIT Press.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59,* 433–460.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.